SEM for those who think they don't care

Vince Wiggins

Vice President, Scientific Development

StataCorp LP

2011 London Stata Users Group Meeting

SEM —The short story

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \Gamma\boldsymbol{x} + \alpha + \zeta$$

Where:

- $\boldsymbol{y}$, $\boldsymbol{x}$, α and ζ are vector
- $\boldsymbol{y}$ and $\boldsymbol{x}$ may contain both latent and observed variables
- ζ is a vector of errors
- $Cov(\boldsymbol{X}, \zeta) = 0$

SEM —The short story

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \Gamma\boldsymbol{x} + \alpha + \zeta$$

Where:

- $\boldsymbol{y}$, $\boldsymbol{x}$, α and ζ are vector
- $\boldsymbol{y}$ and $\boldsymbol{x}$ may contain both latent and observed variables
- ζ is a vector of errors
- $Cov(\boldsymbol{X}, \zeta) = 0$

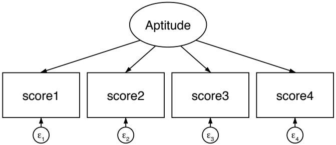
Some interesting things to note:

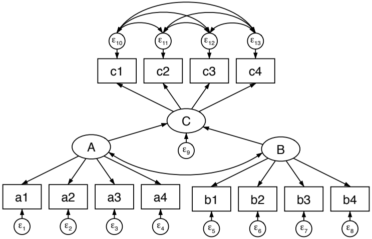
- y's can depend on other y's
- Ignore (mostly) the extensively published rumors that $\boldsymbol{y}$, $\boldsymbol{x}$, and/or ζ must be multivariate normal

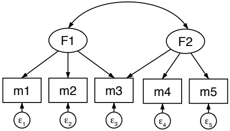
SEM subsumes and extends most linear models.

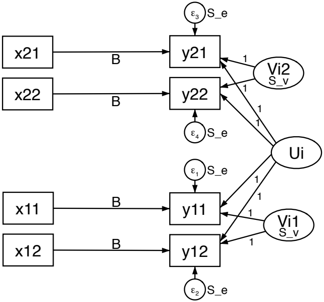
I'm not going to talk about what most SEMers (SEMians?) use SEM for.

Path diagrams

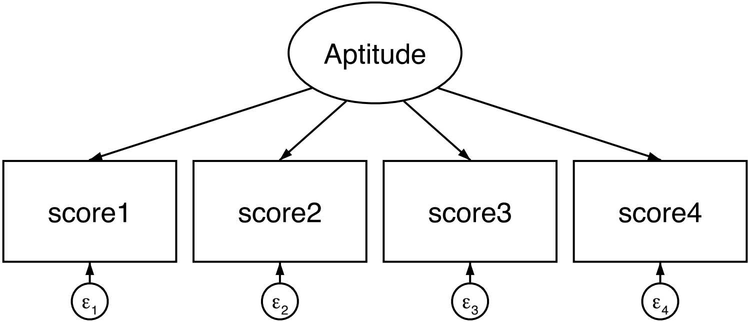




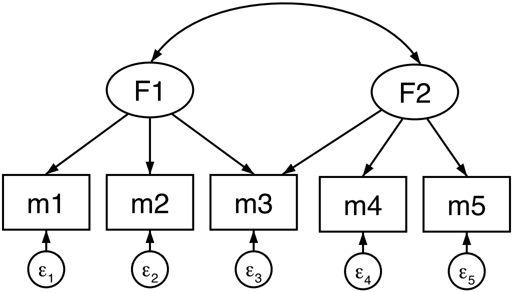




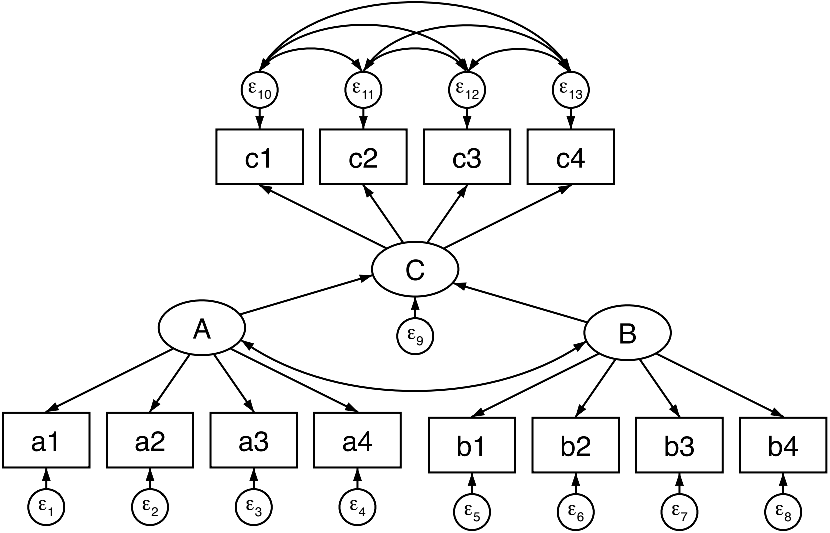
Measurement models/components



Multiple factor models (confirmatory or otherwise)



Full SEM models



I am also not going to talk about

- Extensions to linear and multivariate regression
- Extensions to SURE (including missing values in some *y*'s)
- MIMIC models
- Correlations with missing data
- High-order CFA models
- Correlated uniqueness models
- SEM of latent endogenous variables measured by indicators/measurements

Simultaneous systems and other forms of endogeneity

$$y_1 = \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \epsilon_1$$
$$y_2 = \beta_4 y_1 + \beta_5 x_1 + \beta_6 x_3 + \epsilon_2$$

Simultaneous systems and other forms of endogeneity

$$y_1 = \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \epsilon_1$$
$$y_2 = \beta_4 y_1 + \beta_5 x_1 + \beta_6 x_3 + \epsilon_2$$

```
.  reg3 (y1 y2 x1 x2) (y2 y1 x1 x3)
```

Simultaneous systems and other forms of endogeneity

$$y_1 = \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \epsilon_1$$
$$y_2 = \beta_4 y_1 + \beta_5 x_1 + \beta_6 x_3 + \epsilon_2$$

```
.  reg3 (y1 y2 x1 x2) (y2 y1 x1 x3)
.  sem (y1 <- y2 x1 x2) (y2 <- y1 x1 x3), cov(e.y1*e.y1)
```

Simultaneous systems and other forms of endogeneity

$$y_1 = \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \epsilon_1$$
$$y_2 = \beta_4 y_1 + \beta_5 x_1 + \beta_6 x_3 + \epsilon_2$$

```
. reg3 (y1 y2 x1 x2) (y2 y1 x1 x3)
. sem (y1 <- y2 x1 x2) (y2 <- y1 x1 x3), cov(e.y1*e.y1)
```

Simultaneous systems and other forms of endogeneity

$$y_1 = \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \epsilon_1$$
$$y_2 = \beta_4 y_1 + \beta_5 x_1 + \beta_6 x_3 + \epsilon_2$$

```
.  reg3 (y1 y2 x1 x2) (y2 y1 x1 x3)
.  sem (y1 <- y2 x1 x2) (y2 <- y1 x1 x3), cov(e.y1*e.y1)
```

SEM extensions

- control and constrain the structure of the error covariance matrix
- Obtain SEs, confidence intervals (CIs), etc. that are robust to lack of independence groups of observations —option **`vce(cluster <group>)`**.
- Handle missing data in the dependent variables, so long as it is missing on observables.
- Estimate via GMM (generalized method of moments) —option **`method(adf)`**.
- Estimate direct, indirect, and total effects of all regressors, including the y's —**`estat teffects`**

Multilevel random effects models

$$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$$

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\quad I = 3, J = 2, K = 2$

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\qquad I = 3, J = 2, K = 2$

```
.   set obs 3
```

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\quad I = 3, J = 2, K = 2$

```
.   gen i = _n
```

i
1
2
3

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $I = 3, J = 2, K = 2$

```
.   gen Ui = rnormal()
```

i	Ui
1	μ_1
2	μ_2
3	μ_3

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $I = 3, J = 2, K = 2$

```
.   expand 2
```

i	Ui
1	μ_1
2	μ_2
3	μ_3
1	μ_1
2	μ_2
3	μ_3

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\qquad I = 3, J = 2, K = 2$

```
. by i, sort: gen j = _n
```

i	Ui	j
1	μ_1	1
2	μ_2	1
3	μ_3	1
1	μ_1	2
2	μ_2	2
3	μ_3	2

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\qquad I = 3, J = 2, K = 2$

```
. gen Vij = rnormal()
```

i	Ui	j	Vij
1	μ_1	1	ν_1
2	μ_2	1	ν_2
3	μ_3	1	ν_3
1	μ_1	2	ν_4
2	μ_2	2	ν_5
3	μ_3	2	ν_6

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $I = 3, J = 2, K = 2$

```
. expand 2
```

i	Ui	j	Vij
1	μ_1	1	ν_1
2	μ_2	1	ν_2
3	μ_3	1	ν_3
1	μ_1	2	ν_4
2	μ_2	2	ν_5
3	μ_3	2	ν_6
1	μ_1	1	ν_1
2	μ_2	1	ν_2
3	μ_3	1	ν_3
1	μ_1	2	ν_4
2	μ_2	2	ν_5
3	μ_3	2	ν_6

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk} \qquad I = 3, J = 2, K = 2$

```
. by i j, sort: gen k = _n
```

i	Ui	j	Vij	k
1	μ_1	1	ν_1	1
2	μ_2	1	ν_2	1
3	μ_3	1	ν_3	1
1	μ_1	2	ν_4	1
2	μ_2	2	ν_5	1
3	μ_3	2	ν_6	1
1	μ_1	1	ν_1	2
2	μ_2	1	ν_2	2
3	μ_3	1	ν_3	2
1	μ_1	2	ν_4	2
2	μ_2	2	ν_5	2
3	μ_3	2	ν_6	2

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\qquad I = 3, J = 2, K = 2$

```
. gen Eijk = rnormal()
```

i	Ui	j	Vij	k	Eijk
1	μ_1	1	ν_1	1	ϵ_1
2	μ_2	1	ν_2	1	ϵ_2
3	μ_3	1	ν_3	1	ϵ_3
1	μ_1	2	ν_4	1	ϵ_4
2	μ_2	2	ν_5	1	ϵ_5
3	μ_3	2	ν_6	1	ϵ_6
1	μ_1	1	ν_1	2	ϵ_7
2	μ_2	1	ν_2	2	ϵ_8
3	μ_3	1	ν_3	2	ϵ_9
1	μ_1	2	ν_4	2	ϵ_{10}
2	μ_2	2	ν_5	2	ϵ_{11}
3	μ_3	2	ν_6	2	ϵ_{12}

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\quad I = 3, J = 2, K = 2$

```
. gen x = uniform()
```

i	Ui	j	Vij	k	Eijk	x
1	μ_1	1	ν_1	1	ϵ_1	x_1
2	μ_2	1	ν_2	1	ϵ_2	x_2
3	μ_3	1	ν_3	1	ϵ_3	x_3
1	μ_1	2	ν_4	1	ϵ_4	x_4
2	μ_2	2	ν_5	1	ϵ_5	x_5
3	μ_3	2	ν_6	1	ϵ_6	x_6
1	μ_1	1	ν_1	2	ϵ_7	x_7
2	μ_2	1	ν_2	2	ϵ_8	x_8
3	μ_3	1	ν_3	2	ϵ_9	x_9
1	μ_1	2	ν_4	2	ϵ_{10}	x_{10}
2	μ_2	2	ν_5	2	ϵ_{11}	x_{11}
3	μ_3	2	ν_6	2	ϵ_{12}	x_{12}

Simulating multilevel random effects

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $I = 3, J = 2, K = 2$

```
. gen y = x + Ui + Vij + Eijk
```

i	Ui	j	Vij	k	Eijk	x	y
1	μ_1	1	ν_1	1	ϵ_1	x_1	y_1
2	μ_2	1	ν_2	1	ϵ_2	x_2	y_2
3	μ_3	1	ν_3	1	ϵ_3	x_3	y_3
1	μ_1	2	ν_4	1	ϵ_4	x_4	y_4
2	μ_2	2	ν_5	1	ϵ_5	x_5	y_5
3	μ_3	2	ν_6	1	ϵ_6	x_6	y_6
1	μ_1	1	ν_1	2	ϵ_7	x_7	y_7
2	μ_2	1	ν_2	2	ϵ_8	x_8	y_8
3	μ_3	1	ν_3	2	ϵ_9	x_9	y_9
1	μ_1	2	ν_4	2	ϵ_{10}	x_{10}	y_{10}
2	μ_2	2	ν_5	2	ϵ_{11}	x_{11}	y_{11}
3	μ_3	2	ν_6	2	ϵ_{12}	x_{12}	y_{12}

Sorting by groups

$$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk} \qquad I = 3, J = 2, K = 2$$

i	Ui	j	Vij	k	Eijk	x	y
1	μ_1	1	ν_1	1	ϵ_1	x_1	y_1
1	μ_1	1	ν_1	2	ϵ_2	x_2	y_2
1	μ_1	2	ν_2	1	ϵ_3	x_3	y_3
1	μ_1	2	ν_2	2	ϵ_4	x_4	y_4
2	μ_2	1	ν_3	1	ϵ_5	x_5	y_5
2	μ_2	1	ν_3	2	ϵ_6	x_6	y_6
2	μ_2	2	ν_4	1	ϵ_7	x_7	y_7
2	μ_2	2	ν_4	2	ϵ_8	x_8	y_8
3	μ_3	1	ν_5	1	ϵ_9	x_9	y_9
3	μ_3	1	ν_5	2	ϵ_{10}	x_{10}	y_{10}
3	μ_3	2	ν_6	1	ϵ_{11}	x_{11}	y_{11}
3	μ_3	2	ν_6	2	ϵ_{12}	x_{12}	y_{12}

Reshape 1

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk} \qquad I = 3, J = 2, K = 2$

```
. egen ij = group(i j)
. reshape wide eijk y x, i(ij) j(k)
```

				$k=1$			$k=2$		
i	**Ui**	**j**	**Vij**	**Eij1**	**x1**	**y1**	**Eij2**	**x2**	**y2**
1	μ_1	1	ν_1	ϵ_1	x_1	y_1	ϵ_2	x_2	y_2
1	μ_1	2	ν_2	ϵ_3	x_3	y_3	ϵ_4	x_4	y_4
2	μ_2	1	ν_1	ϵ_5	x_5	y_5	ϵ_6	x_6	y_6
2	μ_2	2	ν_2	ϵ_7	x_7	y_7	ϵ_8	x_8	y_8
3	μ_3	1	ν_1	ϵ_9	x_9	y_9	ϵ_{10}	x_{10}	y_{10}
3	μ_3	2	ν_2	ϵ_{11}	x_{11}	y_{11}	ϵ_{12}	x_{12}	y_{12}

Variable names above are not quite what `reshape` gives.

Reshape 2

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk} \qquad I = 3, J = 2, K = 2$

```
. drop ij
. reshape wide eij1 eij2 y1 y1 x1 x2, i(i) j(k)
```

		$j=1$								$j=2$							
		$k=1$				$k=2$				$k=1$				$k=2$			
i	Ui	$Vi1$	$Ei11$	$x11$	$y11$	$Vi1$	$Ei12$	$x12$	$y12$	$Vi2$	$Ei21$	$x21$	$y21$	$Vi2$	$Ei22$	$x22$	$y22$
1	μ_1	ν_1	ϵ_1	x_1	y_1	ν_1	ϵ_2	x_2	y_2	ν_2	ϵ_3	x_3	y_3	ν_2	ϵ_4	x_4	y_4
2	μ_2	ν_1	ϵ_5	x_5	y_5	ν_1	ϵ_6	x_6	y_6	ν_2	ϵ_7	x_7	y_7	ν_2	ϵ_8	x_8	y_8
3	μ_3	ν_1	ϵ_9	x_9	y_9	ν_1	ϵ_{10}	x_{10}	y_{10}	ν_2	ϵ_{11}	x_{11}	y_{11}	ν_2	ϵ_{12}	x_{12}	y_{12}

Reshape 2

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $I = 3, J = 2, K = 2$

```
. drop ij
. reshape wide eij1 eij2 y1 y1 x1 x2, i(i) j(k)
```

		$j=1$								$j=2$							
		$k=1$				$k=2$				$k=1$				$k=2$			
i	Ui	$Vi1$	$Ei11$	$x11$	$y11$	$Vi1$	$Ei12$	$x12$	$y12$	$Vi2$	$Ei21$	$x21$	$y21$	$Vi2$	$Ei22$	$x22$	$y22$
1	μ_1	ν_1	ϵ_1	x_1	y_1	ν_1	ϵ_2	x_2	y_2	ν_2	ϵ_3	x_3	y_3	ν_2	ϵ_4	x_4	y_4
2	μ_2	ν_1	ϵ_5	x_5	y_5	ν_1	ϵ_6	x_6	y_6	ν_2	ϵ_7	x_7	y_7	ν_2	ϵ_8	x_8	y_8
3	μ_3	ν_1	ϵ_9	x_9	y_9	ν_1	ϵ_{10}	x_{10}	y_{10}	ν_2	ϵ_{11}	x_{11}	y_{11}	ν_2	ϵ_{12}	x_{12}	y_{12}

Think of each bounded column set as a linear regression.

Reshape 2

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\quad I = 3, J = 2, K = 2$

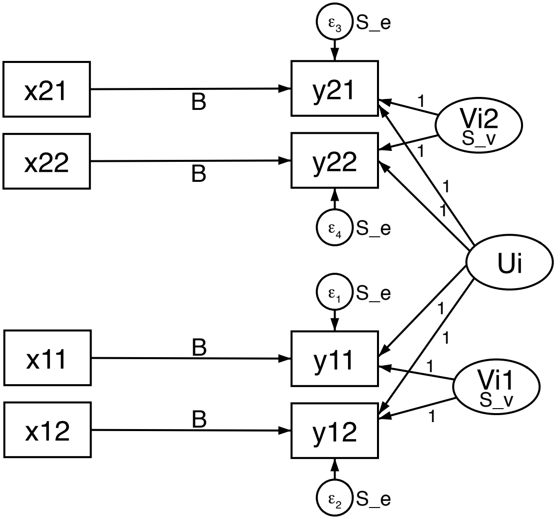
```
. drop ij
. reshape wide eij1 eij2 y1 y1 x1 x2, i(i) j(k)
```

		$j=1$								$j=2$							
		$k=1$				$k=2$				$k=1$				$k=2$			
i	Ui	$Vi1$	$Ei11$	$x11$	$y11$	$Vi1$	$Ei12$	$x12$	$y12$	$Vi2$	$Ei21$	$x21$	$y21$	$Vi2$	$Ei22$	$x22$	$y22$
1	μ_1	ν_1	ϵ_1	x_1	y_1	ν_1	ϵ_2	x_2	y_2	ν_2	ϵ_3	x_3	y_3	ν_2	ϵ_4	x_4	y_4
2	μ_2	ν_1	ϵ_5	x_5	y_5	ν_1	ϵ_6	x_6	y_6	ν_2	ϵ_7	x_7	y_7	ν_2	ϵ_8	x_8	y_8
3	μ_3	ν_1	ϵ_9	x_9	y_9	ν_1	ϵ_{10}	x_{10}	y_{10}	ν_2	ϵ_{11}	x_{11}	y_{11}	ν_2	ϵ_{12}	x_{12}	y_{12}

Think of each bounded column set as a linear regression.

With some creative constraints and a seemingly unrelated regressions structure, this is the estimator for a multilevel random-effects model.

Path diagram for multilevel RE model



Likelihood is identical to `xtmixed`

Unbalanced data

- Different number of observations in some groups?

Unbalanced data

- Different number of observations in some groups?
- No worries?

Unbalanced data

- Different number of observations in some groups?
- No worries?
- add `method(mlmv)`

Results in the exactly same estimator as `xtmixed` with unbalanced panels

Unbalanced

$y_{ijk} = \beta x_{ijk} + \mu_i + \nu_{ij} + \epsilon_{ijk}$ $\qquad I = 3, J = 2, K = 2$

				$k = 1$			$k = 2$		
i	**Ui**	**j**	**Vij**	**Eij1**	**x1**	**y1**	**Eij2**	**x2**	**y2**
1	μ_1	1	ν_1	ϵ_1	x_1	y_1	ϵ_2	x_2	y_2
1	μ_1	2	ν_2	ϵ_3	x_3	y_3	ϵ_4	x_4	y_4
2	μ_2	1	ν_1	ϵ_5	x_5	y_5	ϵ_6	x_6	y_6
2	μ_2	2	ν_2	ϵ_7	x_7	y_7	ϵ_8	x_8	y_8
3	μ_3	1	ν_1	ϵ_9	x_9	y_9	ϵ_{10}	x_{10}	y_{10}
3	μ_3	2	ν_2	ϵ_{11}	x_{11}	y_{11}	ϵ_{12}	x_{12}	y_{12}

I should also mention

For the multilevel RE model (and all the other models) SEM supports:

- robust and cluster-robust SEs
- estimation by GMM
- survey data
- missing data – MAR
- heteroskedastic effects at any level
- correlated effects at any level

Uninterested in SEM?

So was I.
I'm interested now.